MaxEnt power spectrum estimation using the Fourier transform for irregularly sampled data applied to a record of stellar luminosity

Robert W. Johnson

Abstract The principle of maximum entropy is applied to the spectral analysis of a data signal with general variance matrix and containing gaps in the record. The role of the entropic regularizer is to prevent one from overestimating structure in the spectrum when faced with imperfect data. Several arguments are presented suggesting that the arbitrary prefactor should not be introduced to the entropy term. The introduction of that factor is not required when a continuous Poisson distribution is used for the amplitude coefficients. We compare the formalism for when the variance of the data is known explicitly to that for when the variance is known only to lie in some finite range. The result of including the entropic measure factor is to suggest a spectrum consistent with the variance of the data which has less structure than that given by the forward transform. An application of the methodology to example data is demonstrated.

Keywords Fourier transform – Power spectral density – Irregular sampling – Maximum entropy data analysis

1 Introduction

The analysis of imperfect data is a common task in science. Given a set of measurements sampled over time, one commonly uses the Fourier transform to estimate the power carried by the signal as a function of frequency. The forward transform is commonly viewed as the best estimate of the amplitude and phase associated with basis functions of independent frequency; however, the indiscriminate use of the forward transform is not appropriate when the data are known to be subject to measurement error, and the problem of irregular sampling is often addressed by *ad hoc* methods of varying subjectivity, such as interpolation (Cenker et al. 1991; Malik et al. 2005) or zero-padding (Boyd 1992).

Bayesian statistical inference is a well-established methodology for dealing with imperfect data (Bretthorst 1988; Sivia 1996; Gregory 2005). The parameters of interest, here the amplitude and phase comprising the power spectral density, are related to the data through a model function which may be nonlinear. When that function is invertible, its inverse is usually called the forward transform of the data, but the methodology applies as well to model functions which are not invertible. The most likely values for the parameters are given by those which maximize their joint distribution, which takes into account both their likelihood as measured by the discrepancy between the model and the data and their possibly non-uniform prior distribution. A non-uniform prior commonly represents the invariant Haar measure under a change of variables, and when the number of parameters exceeds the number of data, a prior based on entropic arguments is often employed. These methods generally fall under the rubric of MaxEnt data analysis, as the optimum of the joint distribution minimizes the residual while simultaneously maximizing the entropy distribution. Essentially, we are extending the Lomb-Scargle method (Lomb 1976; Scargle 1982) to incorporate the effect of the measurement errors on the estimate of the most likely amplitude and phase coefficients.

After a brief review of Bayesian data analysis, we investigate the details of its application to power spectral density estimation using Fourier basis functions. The problem of missing data values for otherwise regular sampling is easily addressed by working with basis functions defined only at the measurement times. We will compare the methodology for the case when the variance matrix of the data is known explicitly to that

Robert W. Johnson

Alphawave Research, Atlanta, GA 30238, USA

for the more likely occurrence of when the variance is known only to lie in some finite range with assumed independent measurements. The primary result is that the entropic prior flattens the power spectrum relative to that produced by the forward transform, which maximizes solely the likelihood distribution. We demonstrate an application of the method to a signal derived from stellar observation, and we close by discussing recent ideas for improvement of the method so that the variance of the coefficients may also be evaluated.

2 Bayesian primer

The mathematical language of Bayesian data analysis is that of conditional probability distribution functions (Durrett 1994; Sivia 1996). We notate "the probability of A given B under conditions C" as

$$\mathrm{prob}(A|B;C) \equiv p(A|_C B) \equiv p^A_B \ , \tag{1}$$

dropping the conditioning statement C when it is unchanging, but its presence is always implied. The sum and product rules of probability theory give rise to the expressions for marginalization and Bayes' theorem,

$$p^A = \int_{\{B\}} p^{A,B}\, \mathrm{d}B \ , \tag{2}$$

$$p^A_B p^B = p^B_A p^A \ , \tag{3}$$

where marginalization follows from the normalization requirement and Bayes' theorem follows from requiring logical consistency of the joint distribution $p^{A,B} = p^{B,A}$. Translated to data analysis, Bayes' theorem relates the evidence given data $\mathbf{y}$ for the parameters $\mathbf{X}$ yielding model $\mathbf{x} = \mathbf{x}(\mathbf{X})$, denoted $p^{\mathbf{X}}_{\mathbf{y}}$, to the likelihood of the data given the parameters $p^{\mathbf{y}}_{\mathbf{X}}$ times the prior distribution for the parameters $p^{\mathbf{X}}$,

$$p^{\mathbf{X}}_{\mathbf{y}} \propto p^{\mathbf{y}}_{\mathbf{X}} p^{\mathbf{X}} \ , \tag{4}$$

where the constant of proportionality $p^{\mathbf{y}}$ represents the chance of measuring the data which in practice is recovered from the normalization requirement of the joint distribution.

The essential feature of Bayesian data analysis which takes it beyond simple least-squares fitting is the use of a non-uniform prior in appropriate circumstances (D'Agostini 1998). The role of the prior is to prevent one from overestimating structure in the model not supported by imperfect data. A prior which appears non-uniform in one's chosen variable generally represents a prior which is uniform under a change of variable to that with invariant Haar measure. For example, consider fitting a two parameter model for a straight line $\mathbf{x} = b\,\mathbf{t} + a$ to a set of data $(\mathbf{t}, \mathbf{y})$ with finite σ_y and $\sigma_t = 0$. A maximum likelihood analysis (or least-squares for independent data) with a prior uniform on both a and b actually has a preferential bias for extreme values of the slope, as seen by transforming that distribution $p^b \propto 1$ to the variable for the angle $\tan\theta = b$. That transformation is given by $p^\theta = p^b\,|db/d\theta|$ such that $p^\theta \propto 1 + \tan^2\theta$. A prior which instead is uniform over the angle $p^\theta \propto 1$ leads to a Cauchy distribution for the slope $p^b \propto (1+b^2)^{-1}$. These priors are compared in Figure 1 panels (a) and (b). Centering the abscissa $\mathbf{t} \to \mathbf{t} - t_0$ for $t_0 \equiv \sum_d t_d \sigma_d^{-2} / \sum_d \sigma_d^{-2}$ yields independent estimates for the slope and intercept $\sigma^2_{ab} = 0$, and the prior for a is uniform $p^a \propto 1$. At t_0, the intercept a is the estimate of the mean of $\mathbf{y}$, through which the line of best fit must pass.

The effect of a non-uniform prior is to shift the location of the solution with maximum evidence away from that given solely by the likelihood factor. Writing the merit function $F = -\log p^{\mathbf{X}}_{\mathbf{y}}$ as a sum of residual and prior terms $F = R + P$ and dropping the term for the normalization constant, a non-uniform prior $\nabla P \neq 0$ provides an optimal solution $\nabla F(\mathbf{X}_F) = 0$ when the data have very little to say $\nabla R \approx 0$. Returning to our example, let us consider a set of N_d independent measurements with uniform variance and express the

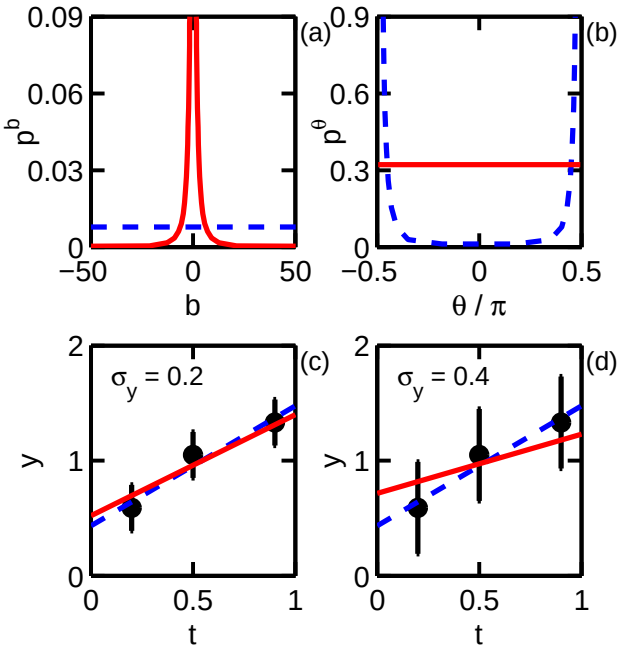

Fig. 1 Prior distributions (a) and (b) for the slope $b = \tan\theta$ of a linear fit. The dashed line shows a prior uniform in b while the solid line shows a prior uniform in θ. The model function $\mathbf{x}_F$ (solid) has a slope with less magnitude than that of $\mathbf{x}_R$ (dashed) which decreases as the variance increases from (c) to (d)

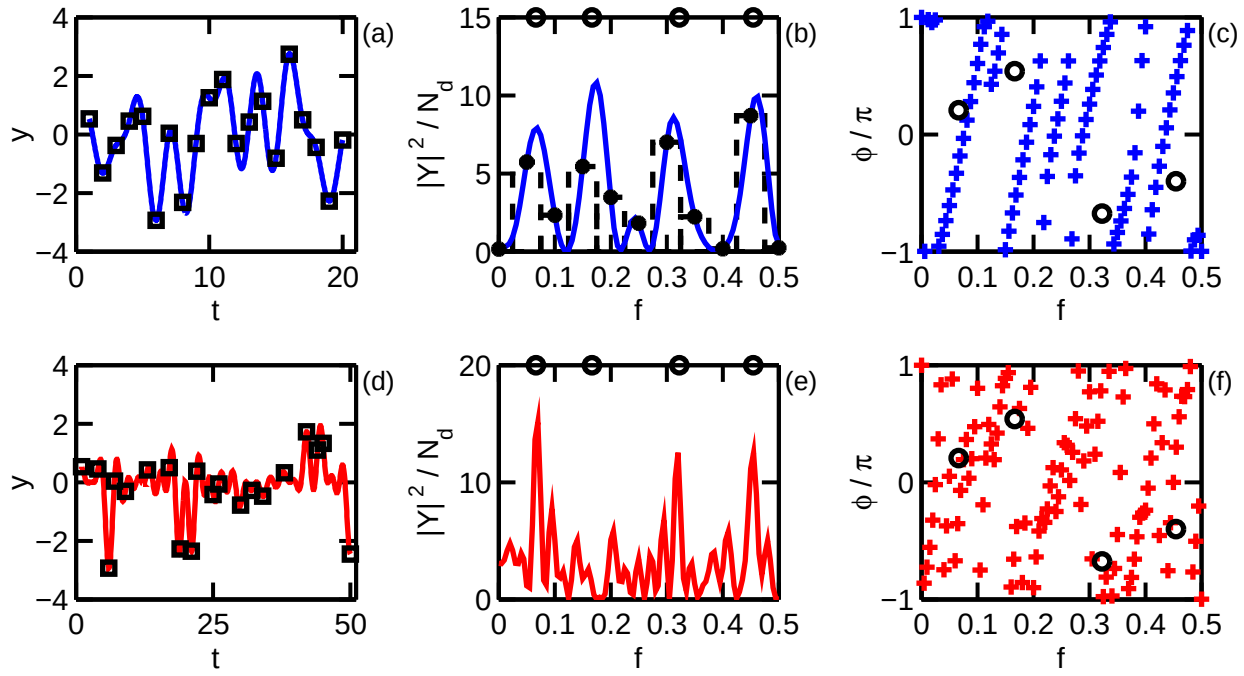

Fig. 2 A regularly sampled signal comprised of four sinusoids of unit amplitude, displayed as squares in (a), has the power spectrum shown in (b), with the signal frequencies indicated by circles at the top and the discrete Fourier transform periodogram marked by dots with the bin widths as dashed lines, and the phase spectrum is shown in (c) with the values of the signal phases circled. When irregularly sampled with the same number of measurements (d), the power spectrum (e) is able (but not guaranteed) to resolve the signal components. The phase spectrum (f) is not as clearly structured as for regular sampling. The interpolated reconstructions displayed as lines in (a) and (d) reproduce the signal values at the measurement times

residual term $R = r^2/2\sigma_y^2$ for $r^2 \equiv \sum_d (x_d - y_d)^2$. As the data become less reliable $\sigma_y^2 \to \infty$, the prior term dominates the gradient and the estimate for b tends to zero, indicating that the data may be characterized solely by their mean rather than exhibiting a linear relation to the abscissa. We compare the model vectors $\mathbf{x}_R$ and $\mathbf{x}_F$ for a set of three data points at two values of σ_y^2 in Figure 1 panels (c) and (d), where one can see how the slope diminishes as the variance increases. The role of the prior is to prevent one from overestimating structure in the model when working with imperfect data by taking into account the measure factor for the chosen parametrization. In that sense, Bayesian analysis provides the most conservative estimate consistent with the data.

3 One-sided discrete CFT

Let us now briefly review the theory of the continuous Fourier transform (CFT) in terms of its discrete application (dCFT). Suppose there exists a signal $y_u(t_u)$, where the subscript u reminds us that the data are given in terms of unit bearing quantities, sampled at regular intervals $t_u \equiv t\Delta_t$ for integer $t \in [1, N_t]$ and defining the unit of time $\Delta_t \equiv 1$, possibly with missing values $N_d < N_t$ such that $\mathbf{t}$ contains only the measurement times and $\mathbf{y}$ the corresponding values. The critical frequency for aliasing is $f_c \equiv (2\Delta_t)^{-1}$ and relates to the periodicity of the spectrum on an infinite frequency axis. For a real signal $y_u \equiv yu_y$ the amplitude spectrum has conjugate symmetry about zero frequency, and so we can restrict consideration to the positive frequency axis $f_u \equiv f\Delta_f$ for integer $f \in [0, N_f]$ and $\Delta_f \equiv (2N_f\Delta_t)^{-1}$. The Fourier basis functions may be represented as a matrix $\mathbf{\Theta}$, where

$$\Theta_{ft} = \sqrt{2}\exp(i2\pi f_u t_u) = \sqrt{2}\exp(i\pi ft/N_f)\ , \tag{5}$$

so that the forward transform $Y_f = \sum_t \Theta_{ft} y_t \Delta_t$ can be written as a matrix multiplication $\mathbf{Y} = \mathbf{\Theta} \boldsymbol{D}_t \mathbf{y}$, where $\boldsymbol{D}_t = \Delta_t \boldsymbol{I}$ is a diagonal matrix whose entries are all Δ_t. The factor $\sqrt{2}$ accounts for the one-sided nature of the transform, representing the response at negative frequencies which differs only by conjugation. The signal energy defined by the sum of squared data values,

$$E_y \equiv \sum_t y_t^2 \Delta_t = \mathbf{y}^T \boldsymbol{D}_t \mathbf{y}\ , \tag{6}$$

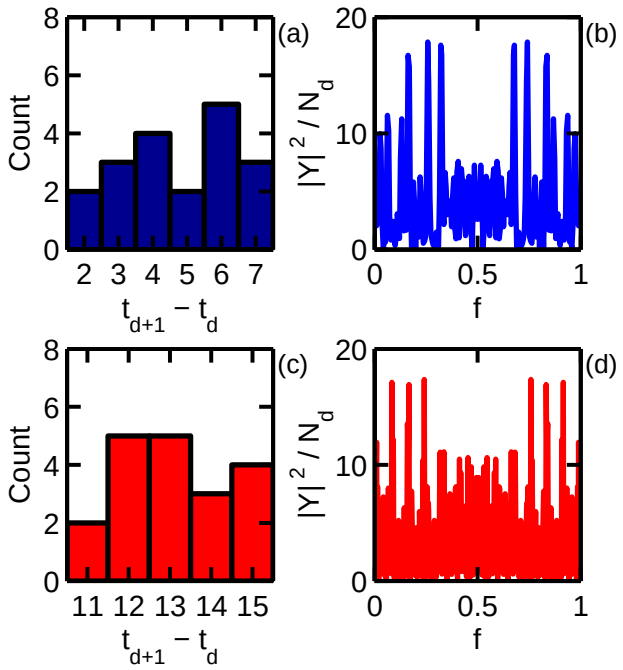

Fig. 3 The critical frequency f_c for irregular sampling depends upon the greatest common factor of the measurement times. In (a) and (c) are histograms of the inter-measurement period for two samplings of the same signal, and their corresponding power spectra in (b) and (d) are symmetric about the critical frequency

is equal to the spectral energy defined in terms of the transform coefficients $E_Y \equiv \mathbf{Y}^\dagger \boldsymbol{D}_f \mathbf{Y}$, where $\boldsymbol{D}_f$ is a diagonal matrix whose entries are Δ_f except for the first and last which are $\Delta_f/2$. The edge-most pixels corresponding to frequencies 0 and N_f have a width only half that of the others when evaluating the integral over the frequency axis $\int_0^{1/2} df_u \rightarrow \sum_f \Delta_f$ in order for the limits to be strictly respected, in contrast to the usual conventions for forming a one-sided power spectral density (psd) from a two-sided Fourier transform which gives those pixels the same weighting as the others. The units of the signal energy differ from those of physical energy by a factor of the load impedance, and the amplitude of the transform coefficients carries units of $u_Y = u_y \Delta_t$, as seen from the inverse transform $\mathbf{y} = \mathrm{Re}\, \boldsymbol{\Theta}^\dagger \boldsymbol{D}_f \mathbf{Y}$, where $\boldsymbol{\Theta}^\dagger \equiv (\boldsymbol{\Theta}^*)^T$ is the conjugate transpose of the basis functions.

In order for the reconstruction to replicate the data (up to round-off errors), one must take a sufficient discretization of the frequency axis. For regularly sampled data $N_d = N_t$, one has the requirement $N_f \geq \lceil N_t/2$, and when N_f is at its minimum the psd of the one-sided dCFT corresponds to that of the discrete Fourier transform periodogram (Press et al. 1992) up to the factors of 2 for the edge-most pixels, as seen in Figure 2 panels (a) through (c). As N_f increases, the resolution along the frequency axis improves so that the remainder of the CFT is evaluated. For irregularly sampled data $N_d < N_t$, one requires $N_f \simeq N_t/2$, but to achieve sufficient resolution it is better to take $N_f = 2N_t$, as shown in Figure 2 panels (d) through (f). Virtually all cases of irregular sampling will correspond to the missing values problem once the greatest common factor of the measurement times is identified as Δ_t, but if the individual measurement durations are not all equal then a suitably generalized $\boldsymbol{D}_t$ must be used. To interpolate (or extrapolate) the data, one simply replaces $\mathbf{t} \rightarrow \mathbf{t}'$ in the inverse transform. One caveat is that the inverse transform is required to agree with the data only at the measurement times, so that for increasing resolution along the frequency axis the interpolant is driven progressively towards the signal mean.

The identification of f_c, which is the lowest positive (nonzero) frequency whose basis function is entirely real over the measurement times, is confirmed by evaluating the psd over the domain $f_u \in [0, 1]$, as seen in Figure 3, where panels (a) and (c) are histograms of the inter-measurement period for two signals and the corresponding spectra in (b) and (d) are symmetric about $f_u = 1/2$, recalling $\Delta_t \equiv 1$. There is no great mystery as to how an irregular sampling can resolve a frequency above the Nyquist limit for the corresponding regular case, as it is the Nyquist limit which must be defined in terms of the irregular sampling. When implementing the continuous Fourier transform in a discrete setting, it is the sampling of the signal which induces the periodicity in the spectrum, while the finite signal duration causes side-lobes to appear in the point spread function (Johnson 2012). The location of the critical frequency f_c must be known so that the normalization of the power spectrum can be evaluated over one (half) period of the frequency axis.

While the amplitude or power spectrum usually receives more attention from investigators, it is actually the phase spectrum which carries most of the information contained in the signal. An amusing example of this phenomenon is the demonstration of how to make a mandrill look like a girl. Taking the 2D Fourier transform of two standard test images shown in Figure 4 panels (a) and (b) using a stock FFT algorithm, one can combine the amplitude spectrum of one image with the phase spectrum of the other to produce two new images from the inverse transform. The resulting images will appear to the eye to resemble the original image associated with the phase used in the combination rather than the amplitude. These two combinations of our test images are displayed in Figure 4 panels (c) and (d), where indeed exchanging the phases has made the mandrill look like a girl and *vice versa*.

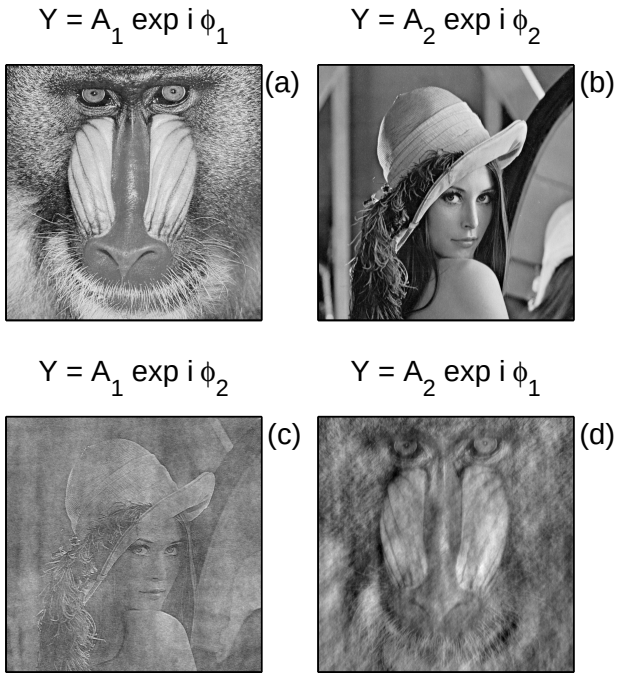

Fig. 4 The phase spectra of two images Mandrill (a) and Lena (b) may be exchanged to produce two new images (c) and (d) which resemble the image associated with the phase more than that associated with the amplitude

4 MaxEnt psd estimation for known data variance

An important feature of Bayesian data analysis is that it allows one formally to address imperfect data as well as to incorporate a prior contribution to the evidence distribution. Rather than assigning errors to the forward transform coefficients based on the data variance, the evidence for a set of coefficients is evaluated from the likelihood of measuring the signal given its variance, and the prior accounts for the measure factor for the chosen parametrization. When there are many more parameters than data values [what Sivia (1996) calls a "free-form" model], an entropic prior is often used (Skilling 1989; Buck and Macaulay 1992); however, its usual expression is derived from an approximation to the Poisson distribution which introduces unnecessarily a factor for the conversion of units. That factor is often said to be a Lagrange multiplier yet is treated as if it were a parameter of the model, which not a consistent approach. With the obvious generalization of the discrete Poisson distribution to a continuum (described in the Appendix), the prior for the amplitude coefficients can be written without reference to an arbitrary unit factor. In this section, we construct the merit function in terms of physically relevant quantities and show its reduction to the usual form as a consequence of normalization. In its physically normalized form, this merit function can be related to those used in statistical physics and lattice gauge theory.

4.1 Model, residual, and constraint

Suppose that we are also given along with $\mathbf{y}$ a variance matrix $\boldsymbol{V}$ whose entries are the covariance of the measurements $V_{jk} \equiv \sigma^2(y_j, y_k)$, from which we form the normalized weight matrix $\boldsymbol{W} \equiv \boldsymbol{V}^{-1}/\mathrm{Tr}\,\boldsymbol{V}^{-1}$ such that $\mathrm{Tr}\,\boldsymbol{W} = 1$. One then defines the residual signal energy $R_E \equiv \mathbf{r}^T\boldsymbol{W}\mathbf{r}N_d\Delta_t$ in terms of the residual vector $\mathbf{r} \equiv \mathbf{x} - \mathbf{y}$, the weight matrix $\boldsymbol{W}$, and the trace of the time metric $\mathrm{Tr}\,\boldsymbol{D}_t = N_d\Delta_t$, (*cf.* the expression for E_y with $\mathbf{y} \to \mathbf{r}$ and $\boldsymbol{D}_t \to \boldsymbol{W}N_d\Delta_t$). The residual vector ultimately is expressed in terms of the model parameters for amplitude and phase $X_f \equiv A_f \exp^i(\phi_f)$, where the notation $\exp^a(x) \equiv (\exp x)^a = e^{ax}$ is similar in spirit to its trigonometric counterpart, through the model function

$$\mathbf{x} \equiv \mathrm{Re}\,\boldsymbol{\Theta}^\dagger\boldsymbol{D}_f\mathbf{X}\,, \tag{7}$$

$$= \Delta_f\sqrt{2}\left[\frac{1}{2}\left(A_0 + A_{N_f}\cos\pi\mathbf{t}\right) + \sum_{f=1}^{N_f-1} A_f\cos\left(\phi_f - \pi f\mathbf{t}/N_f\right)\right]\,, \tag{8}$$

with parameter domains of A_0 and $A_{N_f} \in [-\infty, \infty]$, $A_f \in [0, \infty]$, and ϕ_f periodic in $[-\pi, \pi]$. The frequencies of 0 and f_c must have a phase of 0 or π because their basis functions are entirely real over $\mathbf{t}$. The assignment of uniform priors $p^A \propto 1$ and $p^\phi \propto 1$ to the amplitude and phase parameters corresponds to finding the maximum likelihood solution for the spectrum given the data and its variance.

The likelihood of a signal $\mathbf{y}$ with errors described by a variance matrix $\boldsymbol{V}$ given its "true" value $\mathbf{x}$ is written as a multivariate Gaussian $p^{\mathbf{y}}_{\mathbf{x},\boldsymbol{V}} \propto \exp^{-1}(\chi^2/2)$, where $\chi^2 \equiv \mathbf{r}^T\boldsymbol{V}^{-1}\mathbf{r}$. The subscript E on R_E reminds us that it carries units equal to the signal energy and must be normalized before appearing in the argument of an exponential function. In statistical physics (Wannier 1969), the normalization of the action appearing in the Maxwell distribution is given by the fluctuation energy of the system. Here, let us write the normalization as $\beta_{1/E}R_E$, where $\beta_{1/E} \equiv 1/kT$ is the inverse thermal energy for some generalized temperature T describing the uncertainty to which the measurements are subject. Rearranging factors, we have

$$\beta_{1/E}R_E = (2\beta_{1/E}N_d\Delta_t/\mathrm{Tr}\,\boldsymbol{V}^{-1})(\mathbf{r}^T\boldsymbol{V}^{-1}\mathbf{r}/2) \tag{9}$$

$$\equiv \beta R\,, \tag{10}$$

where $R = \chi^2/2$. We now argue that $\beta = 1$ as follows. Substituting for the thermal energy,

$$\beta = \frac{N_d \Delta_t / \mathrm{Tr}\, \boldsymbol{V}^{-1}}{kT/2} \,, \tag{11}$$

and recalling that $kT/2$ equals the average fluctuation energy per quadratic degree of freedom, we see that the numerator and denominator are equal to $\bar{\sigma}^2 \Delta_t$, where $\bar{\sigma}^2$ is the reciprocal of the average of the eigenvalues of $\boldsymbol{V}^{-1}$. The normalized residual R is seen to be the ratio of the residual signal energy given by the discrepancy of the model to the thermal energy given by the variance of the data. If one were to scale the residual by some arbitrary factor $\beta \neq 1$, that would be tantamount to saying that the experimentalists contributing the measurements have misrepresented the ratio of the units of their signal to its deviation.

By itself, the residual term R is not sufficient to identify uniquely a maximum likelihood solution to the optimization problem. The reason is because the model function Equation (7) is surjective but not injective. What that means is that, for a sufficient discretization of the frequency axis, there exists a continuous family of coefficients $\{\mathbf{X}_R\}$ that can produce a vanishing residual $R(\mathbf{X}_R) = 0$. While the forward transform $\mathbf{Y}(\mathbf{y})$ is one-to-one, the inverse transform $\mathbf{x}(\mathbf{X})$ such that $\mathbf{x} = \mathbf{y}$ is many-to-one; the forward transform coefficients are identified uniquely as the member of $\{\mathbf{X}_R\}$ whose spectral energy equals the signal energy $E_\mathbf{X} = E_\mathbf{y}$. In order that the maximum likelihood solution should equal the forward transform coefficients, the merit function must be supplemented with a term enforcing the constraint.

Using subscripts to indicate which terms are appearing in the merit function, the negative log likelihood of the data, given the constraint on the power spectrum

$$C \equiv \sum_f A_f^2 \Delta_f / E_y - 1 \,, \tag{12}$$

is written $F_{RC} \equiv R + \lambda C$, where λ is a Lagrange multiplier enforcing $C = 0$. The constrained optimum of R coincides with the unconstrained solution of $\nabla_{\lambda,\mathbf{X}} F_{RC} = 0$, which is a saddle point in the space $(\lambda, \mathbf{X})$. The presence of the constraint makes the assignment of errors to the coefficients difficult; not only are there correlations between the amplitudes and phases but also a restriction on the allowed directions the variation in the amplitudes may take. If one amplitude increases, some other must decrease so that the normalization condition is respected. For these reasons, the results appearing henceforth for the psd are understood to be conditioned on satisfaction of the constraint C, and its variance in principle is recoverable from the Hessian of the merit function but not in a form which is practically useful for the chosen parametrization. In a later section we will discuss some recent thoughts on improvements to the method so that an estimate for the variance of the model parameters may be obtained.

4.2 Entropy and the Poisson distribution

The entropic spectral energy for unnormalized distributions (Skilling 1989; Buck and Macaulay 1992; Sivia 1996) is typically written

$$S_E \equiv \sum_f \left[A_f^2 - m_f - A_f^2 \log(A_f^2/m_f) \right] \Delta_f \,, \tag{13}$$

where the sum over f is understood to take into account the frequency metric $\boldsymbol{D}_f$. The factor m_f represents the default model, which in this case is a flat spectrum $m_f \equiv m$ given by the Lebesgue measure with the same energy as the signal, $m = 2E_y\Delta_t$ such that $m \int_0^{1/2} df_u = E_y$. To evaluate the terms of S_E when $A_f \to 0$, one needs to write $0 \log 0 = \log 0^0 = 0$.

To normalize the Lebesgue measure, one divides the entropic spectral energy by the signal energy, letting us write the normalized entropy $S \equiv S_E/E_y$ as

$$S = \sum_f \left\{ A_f^2 \left[1 - \log(A_f^2/2E_y\Delta_t) \right] \Delta_f/E_y \right\} - 1 \,. \tag{14}$$

Under the condition $\sum_f A_f^2 \Delta_f = E_y$ one could manipulate the terms evaluating to a constant, reducing the entropy to the familiar expression $-S_C = \sum_f p_f \log p_f$ for $p_f = A_f^2/2N_f E_y \Delta_t$ and respecting the pixel measure, but the numerical optimization is over the entire space of $\mathbf{X}$ so that only the explicit constant may be dropped for satisfactory convergence of the algorithm.

The quantity of physical relevance is the negentropy, which is the difference between the maximum attainable entropy and that of any particular configuration and so by definition not negative. The equivalent requirement for the entropy expression is that it be nonpositive. For the following examples, let us suppose $E_y = 1$, and consider first a single frequency $N_f = 1$ whose pixel spans the Nyquist interval $\Delta_f = (2\Delta_t)^{-1}$ such that $m = 2$. In Figure 5 panel (a) we compare S_E and S_C for this case, and we see that they do indeed agree at the location which satisfies the constraint, $A = m^{1/2}$, where both expressions equal zero. The reduced entropy S_C, however, takes both positive and negative values over the unconstrained amplitude axis. Replacing the base e of the logarithm with base 2 such that S_C equals the Shannon entropy, as shown in panel (b), the expressions again agree at the value of the constraint, but now the corresponding S_E as well takes on positive values. We are thus led to believe that the

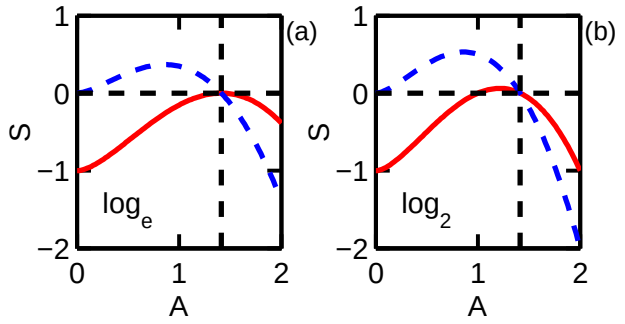

Fig. 5 Comparison of the entropy expressions S_E (solid) and S_C (dashed) for $E_y = 1$ and $N_f = 1$ using logarithm base e in (a) and base 2 in (b), with $A = \sqrt{2}$ and $S = 0$ indicated by dashed lines

physical form of the negentropy expression is that given by S_E in Equation (13) normalized by the Lebesgue measure and with the natural logarithm.

Let us now compare the expressions for S_E and S_C as the number of frequencies spanning the Nyquist interval increases. In Figure 6 we show those expressions as N_f goes from 2 to 5. (We are assuming here a frequency grid dual to the one used elsewhere so that there are no edge effects.) The constraint C is now satisfied not by a single value but by N_f values for A with the required sum of squares. The most obvious difference in behavior is that the maximum of S_C does not remain fixed at $m^{1/2}$ the way that it does for S_E. The presence of both positive and negative values for S_C has a remarkable impact when one follows the mainstream approach of multiplying the entropy by a factor α claimed to play the role of a Lagrange multiplier in the merit function $F_{RC\alpha} \equiv R + \lambda C - \alpha S$ for $S = S_C$. Under enforcement of the constraint $S_C = 0$ for nontrivial amplitudes $\mathbf{A} \neq 0$, contributions with positive entropy must be balanced by contributions with negative entropy, so that the amplitude spectrum is drawn by the residual in either direction away from its nontrivial value where each contribution is zero. The effect is to induce a spikiness to the spectrum, where relatively few large amplitudes with negative entropy (which is unbounded) are balanced by many small amplitudes with positive entropy (which is bounded). While resolution enhancement is often a desired goal (which implies consideration of some point spread function), the use of a term αS_C is not the correct way to go about it. When one uses the normalized negentropy $F_{RC\alpha}$ for $S = S_E/E_y$, the maximum of entropy coincides with the only values which can satisfy the constraint $S_E = 0$, namely the coefficients given by the default model $A_f \equiv m^{1/2}$.

The use of α as a Lagrange multiplier brings up various questions regarding the stopping criterion for the

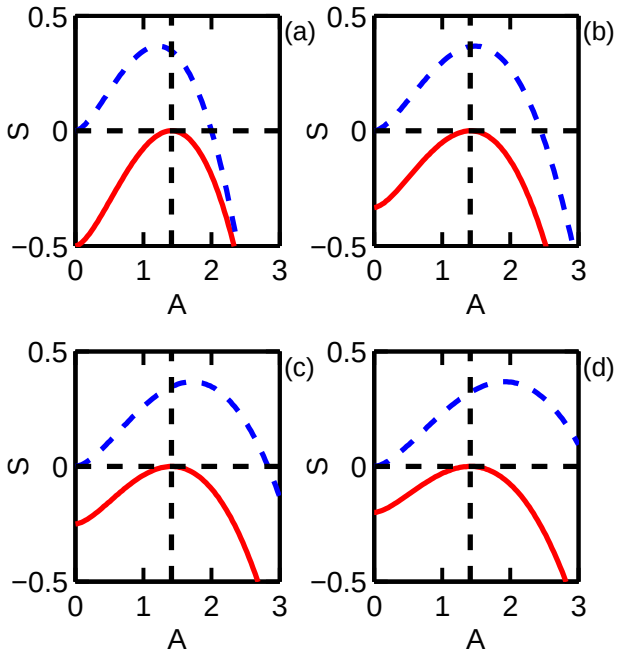

Fig. 6 Comparison of the entropy expressions S_E (solid) and S_C (dashed) for $E_y = 1$ and N_f taking the value 2 in (a), 3 in (b), 4 in (c), and 5 in (d), with the maximum of S_E at $A = \sqrt{2}$ and $S = 0$ indicated by dashed lines

algorithm (Bryan 1990; Strauss et al. 1993; MacKay 1999). Many practitioners use some variation of the method described by Sivia (1996) in which α is treated as a parameter of the model; however, that approach is not consistent with the interpretation of a Lagrange multiplier whose sole purpose is to equate the norms of the gradient vectors for the residual and the constraint. The stationary points of the Lagrangian merit function describe locations satisfying the constraint where the residual changes only in directions which are forbidden. Similarly to the remarks concerning β, the use of $\alpha \neq 1$ is tantamount to an arbitrary rescaling of units between the entropic and signal energies S_E and E_y. By writing $F_{RSC} \equiv R - S + \lambda C$, the MaxEnt solution $\mathbf{X}_F$ is carried smoothly from the forward transform coefficients $\mathbf{X}_{RC} = \mathbf{Y}$ for $\bar{\sigma}^2 \to 0$ to the coefficients with maximum entropy given by the Lebesgue measure as the mean magnitude of the data variance increases, $\mathbf{X}_F \to \mathbf{X}_{SC}$ as $\bar{\sigma}^2 \to \infty$.

Returning to the literature (Skilling 1989; Buck and Macaulay 1992), let us reexamine the arguments leading to the entropy expression found in Equation (13). When invoking the hypothetical troop of monkeys, one supposes not only that the image, here the power spectrum, is comprised of discrete pixels Δ_f but also that the pixel values are themselves described discretely in terms of some presumably small quantum of image, here power,

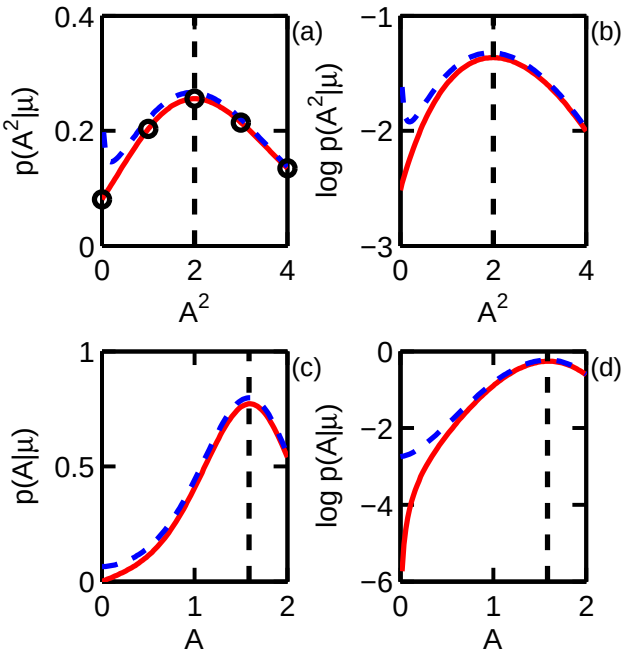

Fig. 7 Comparison of the continuum Poisson distribution (solid) with its Stirling approximation (dashed) for $m = 2$, with the peaks at A_m^2 and A_μ indicated by dashed lines and the corresponding values of the discrete Poisson distribution circled in (a)

such that $A_f^2 = \epsilon\, n_f$ for integer n_f, in order to write the measure as a product of discrete Poisson distributions $p(n_f|\nu_f)$ with parameters $\nu_f \equiv \nu$, taken here to be uniform. Considering a single pixel p_ν^n, the requirement that ϵ be small so that n is large allows the Stirling approximation to the factorial,

$$p_\nu^n = e^{-\nu}\nu^n/n! \quad (15)$$
$$\approx (2\pi n)^{-1/2}\exp(n-\nu-n\log n/\nu)\,, \quad (16)$$

and transforming variables to those for amplitude $a^2 = n$ such that $dn/da = 2a$ lets one write

$$p_\nu^a = p_\nu^n\,|dn/da|\,, \quad (17)$$
$$\approx (2/\pi)^{1/2}\exp(a^2-\nu-a^2\log a^2/\nu)\,. \quad (18)$$

The inverse of the quantum $\epsilon^{-1} \equiv \alpha$ is then identified as the prefactor in the expression $\alpha S_E = \sum_f [a_f^2 - \nu - a_f^2\log(a_f^2/\nu)]\Delta_f$. The motivation for the introduction of α hinges entirely on the discrete form of the Poisson distribution. Its definition here as a unit factor for quantization is not consistent with its interpretation previously as a Lagrange multiplier.

We propose that there is no need to introduce α when one considers the extension of the Poisson distribution to the continuum using the obvious substitution $n! \to \Gamma(n+1)$, where now $n = A^2$ and the parameter in the given units is μ. In the Appendix we discuss the use of this expression as a probability density function. For a single pixel, the measure of the power distribution is given by

$$p(A^2|\mu) = e^{-\mu}\mu^{A^2}/\Gamma(A^2+1)\,, \quad (19)$$

which if maximized at a value $A_m^2 = m$ has a parameter $\mu_m = \exp[\Lambda_1(m+1)]$, using the notation $\Lambda_k(r) \equiv (\partial_r)^k \log\Gamma(r)$ for the polygamma functions with real argument and integer $k \geq 0$. In terms of amplitude, the distribution is

$$p(A|\mu) \propto |A|\,p(A^2|\mu)\,, \quad (20)$$

which has its peak not at A_m but at $A_\mu = \mu^{1/2}$. Using the notation $q \equiv -\log p$, one can write

$$q(A^2|\mu) = \mu - A^2\log\mu + \Lambda_0(A^2+1)\,, \quad (21)$$
$$q(A|\mu) = q(A^2|\mu) - \frac{1}{2}\log A^2\,, \quad (22)$$

having dropped a factor of 2. Under the Stirling approximation,

$$\Lambda_0(A^2+1) \approx A^2\log A^2 - A^2 + \frac{1}{2}\log(2\pi A^2)\,, \quad (23)$$

the expressions for q are

$$q(A^2|\mu) \approx \mu - A^2 + A^2\log(A^2/\mu) + \frac{1}{2}\log A^2\,, \quad (24)$$
$$q(A|\mu) \approx \mu - A^2 + A^2\log(A^2/\mu)\,, \quad (25)$$

having dropped numerical terms. The full expressions for p and q are compared for $m = 2$ in Figure 7, and in panel (a) the corresponding values of the discrete Poisson distribution are circled. Using Equations (21) and (22), we can now write the prior measure for the amplitudes in terms of its contribution P to the merit function $F_{RPC} \equiv R + P + \lambda C$ as

$$P = \sum_f [\Lambda_0(A_f^2+1) - \frac{1}{2}\log A_f^2 - A_f^2\log\mu]\Delta_f/E_y\,, \quad (26)$$

taking into account the pixel measure $\boldsymbol{D}_f$ and dropping the constant term proportional to μ. The parameters A_0 and A_{N_f} have twice the range but only half the pixel width, so their prior normalization is consistent with the others.

4.3 Finding the solution

To solve the optimization problem for $F = F_{RPC}$, one needs the gradient $\mathbf{G} \equiv \nabla F$ and the Hessian $\boldsymbol{H} \equiv \nabla^T\nabla F$, with ∇^T a column vector of derivatives $\nabla^T \equiv [\partial_\lambda, \partial_{\mathbf{X}}]^T$ for $\partial_{\mathbf{X}}$ (a row vector) the covariant gradient

in $\mathbf{X}$, written so that matrix multiplication is embodied in the notation. The gradient of the residual vector $\partial_{\mathbf{X}}\mathbf{r} = \partial_{\mathbf{X}}(\mathbf{x}-\mathbf{y})$ is a matrix with a column index for the parameters $\mathbf{X}$ and a row index for the measurement times $\mathbf{t}$. The Hessian operator has the dyadic form

$$\nabla^T\nabla \equiv \begin{bmatrix} \partial_\lambda \\ \partial_{\mathbf{X}}^T \end{bmatrix} \begin{bmatrix} \partial_\lambda & \partial_{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} \partial_\lambda^2 & \partial_\lambda\partial_{\mathbf{X}} \\ \partial_\lambda\partial_{\mathbf{X}}^T & \partial_{\mathbf{X}}^T\partial_{\mathbf{X}} \end{bmatrix} , \tag{27}$$

and the λ dependence is $\partial_\lambda F = C$, $\partial_\lambda^2 F = 0$, and $\partial_\lambda\partial_{\mathbf{X}}F = \partial_{\mathbf{X}}C$. The residual term has contributions

$$\partial_{\mathbf{X}}R = \mathbf{r}^T\boldsymbol{V}^{-1}\left(\partial_{\mathbf{X}}\mathbf{r}\right) , \tag{28}$$

$$\partial_{\mathbf{X}}^T\partial_{\mathbf{X}}R = \left(\partial_{\mathbf{X}}\mathbf{r}\right)^T\boldsymbol{V}^{-1}\left(\partial_{\mathbf{X}}\mathbf{r}\right) + \mathbf{r}^T\boldsymbol{V}^{-1}\left(\partial_{\mathbf{X}}^T\partial_{\mathbf{X}}\mathbf{r}\right) , \tag{29}$$

where $\left(\partial_{\mathbf{X}}^T\partial_{\mathbf{X}}\mathbf{r}\right)$ is contracted along its time index with $\mathbf{r}^T\boldsymbol{V}^{-1}$. Ordering the parameter vector using the notation $\mathbf{X}^T \equiv [A_0, A_{N_f}, [A_f, \phi_f]_{f=1}^{N_f-1}]$ and letting $\mathbf{u}_f \equiv \pi f\mathbf{t}/N_f$, we can write

$$(\partial_{\mathbf{X}}\mathbf{r})^T = \frac{\Delta_f}{\sqrt{2}}\begin{bmatrix} 1 \\ \cos\left(\pi\mathbf{t}^T\right) \\ \begin{bmatrix} 2\cos\left(\phi_f - \mathbf{u}_f^T\right) \\ -2A_f\sin\left(\phi_f - \mathbf{u}_f^T\right) \end{bmatrix}_f \end{bmatrix} , \tag{30}$$

$$\partial_{\mathbf{X}}^T\partial_{\mathbf{X}}\mathbf{r} = \frac{\Delta_f}{\sqrt{2}}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \begin{bmatrix} 0 & -2\sin\left(\phi_f - \mathbf{u}_f\right) \\ -2\sin\left(\phi_f - \mathbf{u}_f\right) & -2A_f\cos\left(\phi_f - \mathbf{u}_f\right) \end{bmatrix}_f \\ 0 & 0 & \end{bmatrix} , \tag{31}$$

and for $\Psi_\pm(A) \equiv \Lambda_1(A^2) \pm (2A^2)^{-1} - \log\mu + \lambda$, one has

$$\partial_{\mathbf{X}}^T(P+\lambda C) = \frac{\Delta_f}{E_y}\begin{bmatrix} A_0\Psi_+(A_0) \\ A_{N_f}\Psi_+(A_{N_f}) \\ \begin{bmatrix} 2A_f\Psi_+(A_f) \\ 0 \end{bmatrix}_f \end{bmatrix} , \tag{32}$$

$$\partial_{\mathbf{X}}^T\partial_{\mathbf{X}}(P+\lambda C) = \frac{\Delta_f}{E_y}\begin{bmatrix} 2A_0^2\Lambda_2(A_0^2) + \Psi_-(A_0) & 0 & 0 & 0 \\ 0 & 2A_{N_f}^2\Lambda_2(A_{N_f}^2) + \Psi_-(A_{N_f}) & 0 & 0 \\ 0 & 0 & \begin{bmatrix} 4A_f^2\Lambda_2(A_f^2) + 2\Psi_-(A_f) & 0 \\ 0 & 0 \end{bmatrix}_f \\ 0 & 0 & \end{bmatrix} \tag{33}$$

With these evaluations, the solution with maximum evidence $\mathbf{X}_F$ may be found using commonly available numerical optimization routines (Press et al. 1992). Let us compare the amplitude spectra for $\mathbf{Y}$ and $\mathbf{X}_F$ given a signal with unit variance and missing values. The simulated signal displayed in Figure 8 (a) is comprised of four sinusoids of unit amplitude and Gaussian noise of unit variance and has the forward transform power spectral density $|Y_f|^2$ shown in (b) and the phase spectrum shown in (c), where two of the four signal components have been well resolved for this particular sampling. The effect of including the variance is seen in Figure 8 (d), where the peaks have been reduced to a level not far above the noise, and in (e), where the phases have adjusted slightly. Recalling that the psd is proportional to a probability distribution, we see that the evidence for the signal components is reduced by a factor of nearly 2 compared to their likelihood when the signal variance is on the order of the signal magnitude squared. Note that the signal energy of the reconstruction is generally less than the original signal energy, $E_x < E_y$, indicating that phase cancellations occur in the spectrum with maximum evidence.

The effect of the non-uniform prior for the amplitudes A_f has been to draw the power spectrum towards the default model given by a flat spectrum. The amount by which the spectrum is flattened depends upon the magnitude of the variance of the data. The phase parameters ϕ_f are given a uniform prior so that they are free to adjust as needed to bring the model vector $\mathbf{x}$ as close as possible to the data vector $\mathbf{y}$. In practice, the phases are not bound to their principle branch so

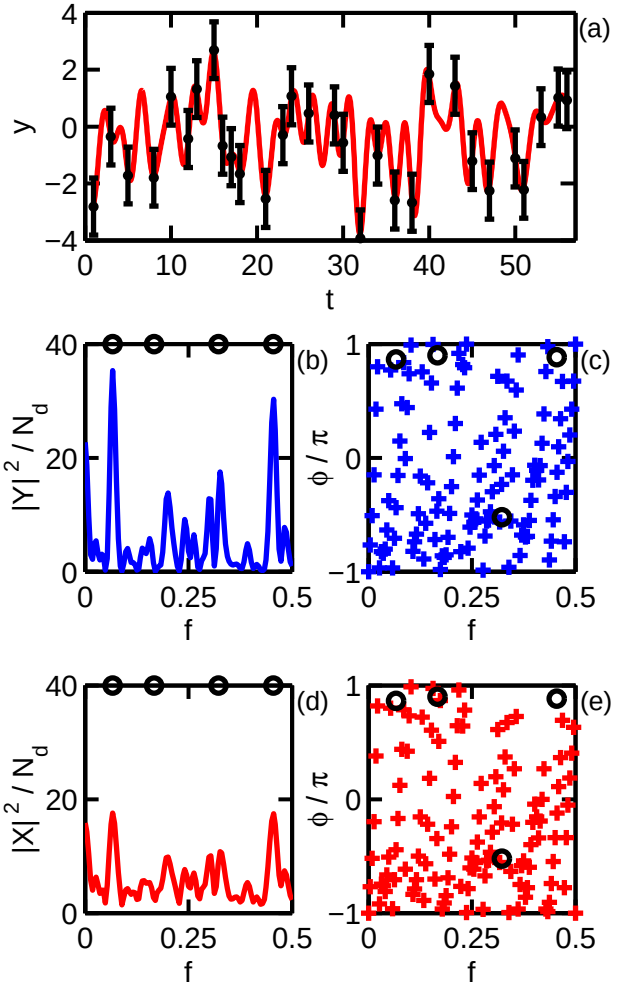

Fig. 8 The irregularly sampled signal is displayed in (a) as dots with error-bars. The maximum likelihood psd in (b) is given by the forward dCFT, and its phases are shown in (c). The maximum evidence psd in (d) accounts for the variance of the data by bringing the psd closer to a flat spectrum, and the phases in (e) have varied slightly. The reconstructed signal is shown in (a) as a solid line

that the algorithm does not get hung up on a branch cut; they are simply reduced to the principle branch after the optimization. The maximum evidence solution is more conservative than that given by the forward transform, in that less structure is assigned to the power distribution. Features which persist in the spectrum are the most likely to be of significance.

5 MaxEnt psd estimation for unknown data variance

Let us now look at how the methodology changes when the variance of the data is known only to lie in some finite range with assumed independent measurements.

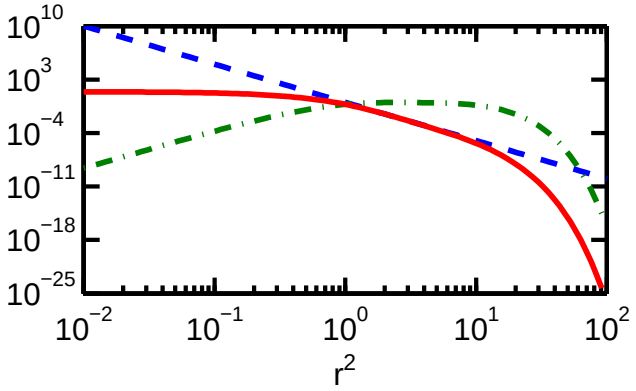

Fig. 9 The likelihood distribution (solid) for data with unknown variance is proportional to the product of two factors, r^{-2a} (dashed) and Δ_Γ (dash-dot), here normalized by $\Gamma(a)$ for $a = 5$ and $\sigma^2 \in [0.1, 1]$

The expressions for P and λC remain the same, so we will focus on the changes to the residual term R. It will prove convenient to work with the squared norm of the residual vector r^2, so that $\nabla r^2/2 = \mathbf{r}^T(\nabla \mathbf{r})$ and $\nabla^T \nabla r^2/2 = (\nabla \mathbf{r})^T(\nabla \mathbf{r}) + \mathbf{r}^T(\nabla^T \nabla \mathbf{r})$. The likelihood distribution is now expressed as an integral over the nuisance parameter for the data deviation σ,

$$\int_{\sigma_0}^{\sigma_1} \sigma^{-N_d-1} \exp\left(\frac{-r^2}{2\sigma^2}\right) d\sigma \propto (r^2)^{-N_d/2} \Delta_\Gamma \ , \qquad (34)$$

where the integrand includes the Jeffreys prior $1/\sigma$ as well as the normalization of the Gaussian and $\Delta_\Gamma \equiv \Gamma(N_d/2, r^2/2\sigma_1^2) - \Gamma(N_d/2, r^2/2\sigma_0^2)$ in terms of the upper incomplete gamma function $\Gamma(a, z) \equiv \int_z^\infty e^{-u} u^{a-1} du$. In the limit of $\sigma_0 \to 0$ and $\sigma_1 \to \infty$, one has $\Delta_\Gamma \to \Gamma(a)$ which is absorbed by the normalization such that the minimum of R is given by $r^2 \to 0$, but for a finite range $\sigma \in [\sigma_0, \sigma_1]$, the divergence of the factor r^{-2a} is canceled by its reciprocal appearing in the continued fraction expression of $\Gamma(a, z) = \exp(-z) z^a (z + \ldots)^{-1}$ so that the likelihood is effectively constant for r^2 below a certain threshold and also reduced significantly for large values of r^2, as seen in Figure 9.

Considering the same signal as in the previous section, let us first suppose that $\sigma_a \in [1, 10]$. The resulting MaxEnt psd is shown in Figure 10 (b), and we see that it is nearly identical to that produced by the previous analysis which assumed a unit variance. When the variance of the data is known only to lie in some finite range, the evidence is dominated by the contribution from the lower bound on the data variance. Let us now suppose an extreme case of $\sigma_b \in [10, 100]$, where the variance exceeds the magnitude of the signal. The psd

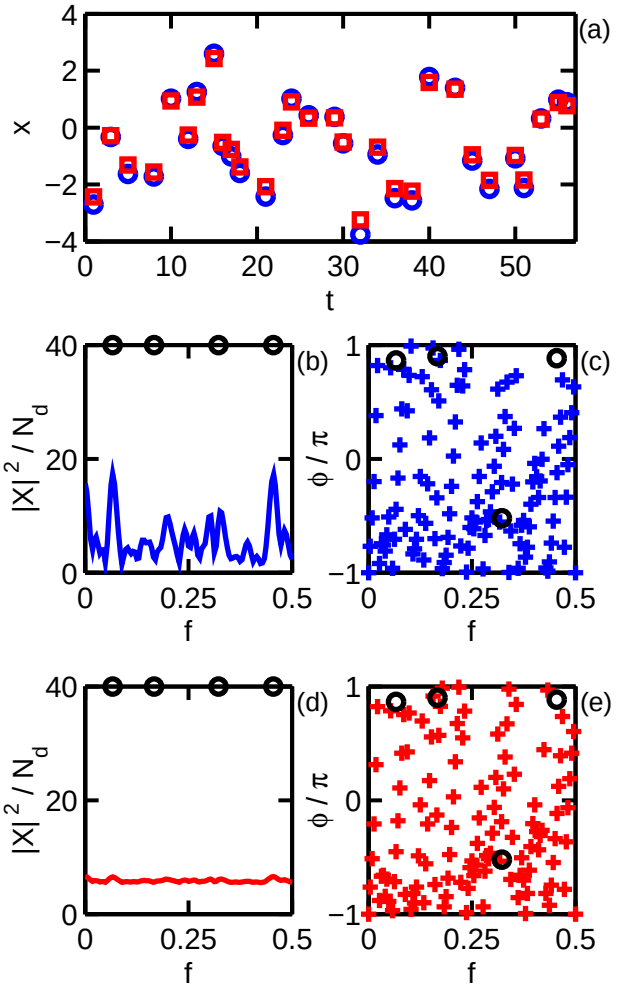

Fig. 10 The MaxEnt psd for unknown data variance in the range $\sigma_a \in [1, 10]$ in (b) and (c) is exceedingly similar to the psd for known variance $\sigma = 1$. When the data is of poor quality, $\sigma_b \in [10, 100]$, the MaxEnt psd in (d) in (e) is very close to the default model of a uniform spectrum. The reconstructions in (a), $\circ$ for σ_a and $\square$ for σ_b, are drawn closer to the signal mean as the lower bound on the data variance increases

in this case, Figure 10 (d), is now very close to the uniform distribution given by the default model of the prior term—the previously resolved signal components register barely a ripple on an otherwise flat power spectrum. Probability theory has prevented us from overestimating structure in the spectrum when the quality of the data is suspect. As the lower bound on the variance increases, the reconstruction from the model function grows closer to the mean of the data, as seen in Figure 10 (a), such that the reconstructed signal's energy decreases from the value given by the original data, $E_b < E_a < E_y$, indicating that phase cancellations become more prominent in the spectrum.

6 Application to a record of stellar luminosity

Let us now apply the MaxEnt methodology to some real data. The signal chosen here is a record of luminosity (Henden 2011) for the star VCas dating from the beginning of January, 2010, to the end of January, 2011. Some measurements are given with a temporal resolution of seconds, but to keep the analysis tractable we assign the data to a daily time axis. When more than one measurement falls on the same day (a rare occurrence) their mean and its variance are used. To reveal the low frequency content, the arithmetic mean is subtracted before conducting the spectral analysis. The choice of the arithmetic rather than the weighted mean is made because it is the arithmetic mean which appears in the 0 frequency bin of the forward transform.

The data span $N_t = 390$ days, as shown in Figure 11 (a), and so a $N_f = 780$ point frequency axis is used. The error bars are hard to see because they are small. The MaxEnt reconstruction also shown in panel (a) is driven towards the signal mean because of the oversampling of the frequency axis. The forward transform psd displayed in panel (b) shows a prominent low frequency peak as well as a few others with a period greater than 30 days. The transform is evaluated on a linear frequency axis but displayed on a logarithmic axis so that the low frequency region is more easily observed. The MaxEnt psd shown in (c) has not changed much, as expected from the small magnitude of the errors, but by incorporating the data variance and the amplitude measure, it provides a more conservative estimate of the spectral density. The phase plots have been suppressed as very little variation is seen between the algorithms for this data set.

The locations of the four lowest frequency peaks are given in Table 1, as is their ratio to the lowest frequency. These frequency peaks do not appear to be in a harmonic relation with the fundamental frequency of the signal. Note that the MaxEnt algorithm does not alter the locations of the peaks but does broaden their widths so that the variance of a frequency estimate increases with the noise level of the data. For this particular data set, a single or few frequency model (Bretthorst 1988) might be more appropriate; however, this signal was chosen simply as an example of the type of data to which the MaxEnt psd algorithm for irregular sampling is applicable.

7 Variance of the psd

In this section we present some recent thoughts on how the methodology might be improved so that an estimate of the variance of the model parameters may be

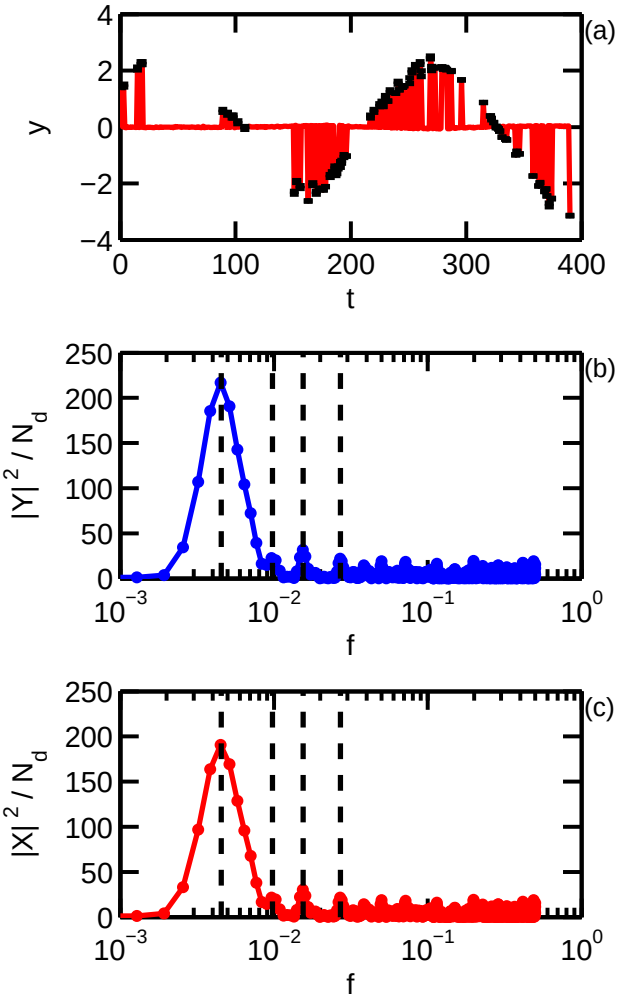

Fig. 11 The MaxEnt reconstruction of the VCas luminosity signal in (a) is driven to the signal mean between the measurement times by the oversampling of the frequency axis. The time axis is given in units of days, and the frequency axis in units of cycles per day. The forward transform psd in (b) displays low frequency structure which becomes more apparent on a logarithmic frequency axis, where the locations of the four lowest frequency peaks are indicated by dashed lines. The MaxEnt psd in (c) is drawn only slightly towards a flat spectrum because of the small magnitude of the data variance. The phase plots have been suppressed as little variation is seen between the forward transform and MaxEnt spectra

obtained. The difficulty with the assignment of errors in the method described so far results from the appearance of the constraint term λC in the Lagrangian, so that the stationary point for the constrained optimum is at a saddle point in the extended parameter space $(\lambda, \mathbf{X})$. Obviously, the solution is to devise a methodology in which no constraint appears, so that the Hessian $\boldsymbol{H}$ of the merit function F at its optimal value $\mathbf{G}(\mathbf{X}_F) = 0$ provides an estimate of the variance of the model parameters. As the normalization of the constraint C is given by the signal energy E_y, that is where we will look for improvements.

Table 1 Peak locations in the power spectral density of the VCas data in units of days

	first	second	third	fourth
period	220.82	102.11	64.67	36.98
freq.	0.004529	0.009793	0.015463	0.027038
ratio	1	2.163	3.415	5.971

Rather than constrain the spectral energy $E_{\mathbf{X}}$ to the value given by the sum of squared data values E_y in Equation (6), what we need to do is evaluate the probability of $E_{\mathbf{X}}$ for any set of coefficients given the data $\mathbf{y}$ and its variance $\boldsymbol{V}$. Let us rewrite the normalized signal energy in terms of the expectation value of the squared data values,

$$E_y/N_d\Delta_t = \langle y_t^2\rangle_{\boldsymbol{W}} \equiv \mathbf{y}^T\boldsymbol{W}\mathbf{y}/\mathrm{Tr}\,\boldsymbol{W}\ , \tag{35}$$

here using the unnormalized weight matrix $\boldsymbol{W} \equiv \boldsymbol{V}^{-1}$. Simply assigning that value a variance given by

$$\langle (y_t^2 - \langle y_t^2\rangle_{\boldsymbol{W}})^2\rangle_{\boldsymbol{W}} = \langle y_t^4\rangle_{\boldsymbol{W}} - \langle y_t^2\rangle_{\boldsymbol{W}}^2 \tag{36}$$

turns out to be a bad idea, as the distribution for the energy $p(E_y|\mathbf{y},\boldsymbol{V})$ is nothing like a Gaussian.

As E_y and $\langle y_t^2\rangle_{\boldsymbol{W}}$ differ only by a scaling factor $N_d\Delta_t$, let us focus our attention on the distribution $p(y^2|\mathbf{y},\boldsymbol{V})$. Considering some signal $\mathbf{y}$, first suppose that its variance is proportional to the identity matrix $\boldsymbol{V} = \sigma^2\boldsymbol{I}$. The individual datum distributions for $k \in [1, N_d]$ are then given by normalized Gaussians,

$$p(y_k|y_t,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp^{-1/2}\left[(y_k - y_t)^2/\sigma^2\right]\ , \tag{37}$$

which need to be folded over to a strictly positive axis $z_k \equiv |y_k|$ by writing

$$p(z_k|y_t,\sigma^2) \propto \exp^{-1/2}\left[(z_k - y_t)^2/\sigma^2\right] \tag{38}$$

$$+\exp^{-1/2}\left[(-z_k - y_t)^2/\sigma^2\right]\ . \tag{39}$$

That distribution is marginalized over k so that

$$p(z|\mathbf{y},\sigma^2) = \sum_k p(z_k|y_t,\sigma^2)/N_d \tag{40}$$

may then be scaled for $y^2 = z^2$ to yield

$$p(y^2|\mathbf{y},\sigma^2) = p(z|\mathbf{y},\sigma^2)|dz/dy^2|\ , \tag{41}$$

which gives the distribution for the estimate of the signal energy given the data and its variance. For the

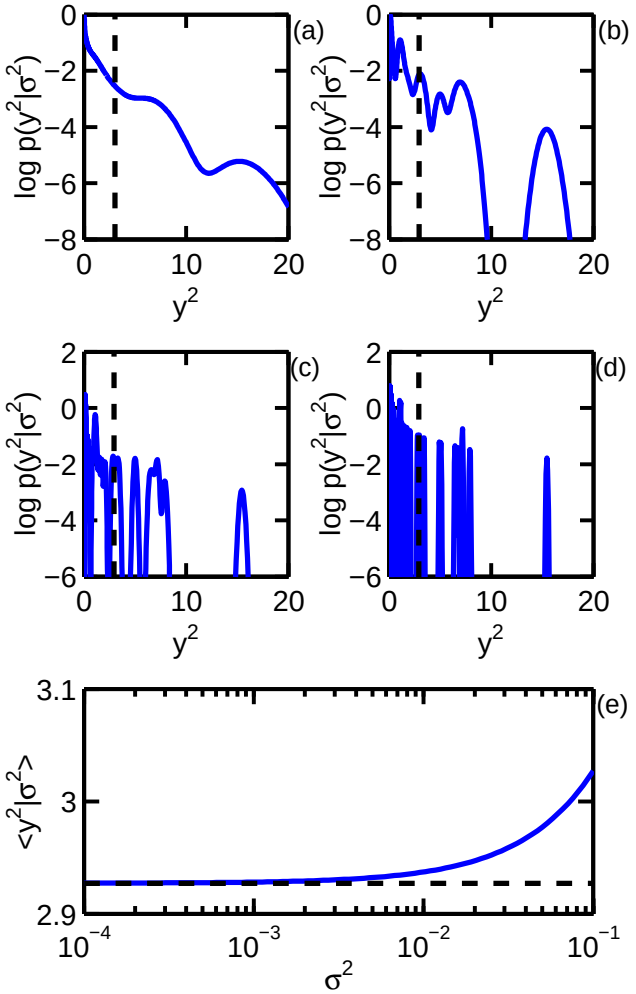

Fig. 12 The logarithm of the distribution $p(y^2|\sigma^2)$ is shown for $\sigma^2 = 10^{-1}$ in (a), 10^{-2} in (b), 10^{-3} in (c), and 10^{-4} in (d), where the expectation value of y^2 is indicated by a dashed line. The expectation value as a function of σ^2 is shown in (e), where the arithmetic mean of the squared data values is indicated by a dashed line

signal $\mathbf{y}$ in Figure 8, let us evaluate $p(y^2|\mathbf{y}, \sigma^2)$ for $\sigma^2 \in [10^{-4}, 10^{-1}]$. In Figure 12 panels (a) through (d) we show the logarithm of that distribution for the four values of σ^2 equal to powers of 10, and in panel (e) we show the expectation value $\int y^2 p(y^2|\mathbf{y}, \sigma^2) dy^2$ as a function of σ^2. As $\sigma^2 \to 0$, the distribution in y^2 approaches a sum of delta distributions located at the values of y_t^2, and the expectation value approaches the arithmetic mean of the squared data values.

If the variance matrix for the data is not proportional to the identity, one has to account for the covariance (off-diagonal elements) when evaluating the distribution for the signal energy. To orthogonalize the data, one must decompose the weight matrix $\boldsymbol{W}$ into

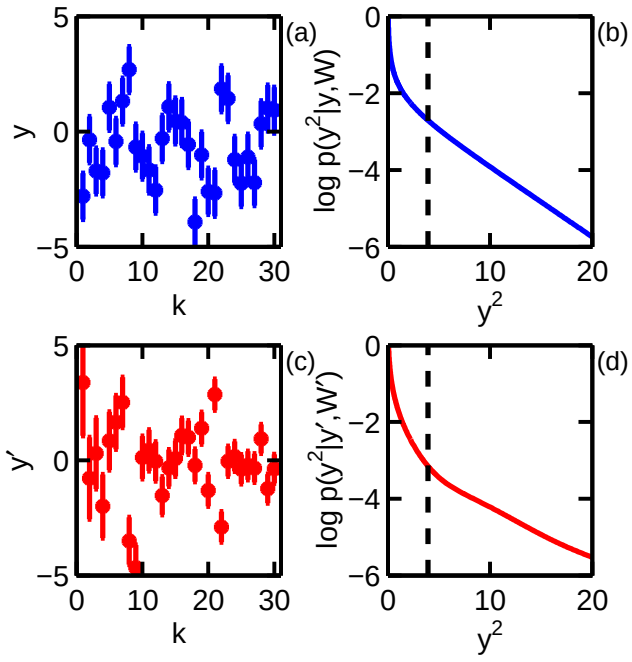

Fig. 13 The datum distributions indexed by k in panel (a) with independent unit variance have a distribution for normalized energy shown in (b). Assuming the data variance is given by a symmetric Toeplitz matrix with the same trace, the orthogonalized data in (c) have the distribution for normalized energy shown in (d), which has the same expectation value for y^2

its eigenvalues $\boldsymbol{W}'$ and eigenvectors $\boldsymbol{E}$, so that

$$\mathbf{y}^T \boldsymbol{W} \mathbf{y} \to \mathbf{y}^T \boldsymbol{E}^T \boldsymbol{W}' \boldsymbol{E} \mathbf{y} \equiv \mathbf{y}'^T \boldsymbol{W}' \mathbf{y}' \tag{42}$$

yields the same expectation value for y^2. The orthogonalized datum distributions

$$p(y_k|y'_t, w'_k) = (2\pi/w'_k)^{-1/2} \exp^{-1/2} \left[(y_k - y'_t)^2 w'_k\right] \ , \tag{43}$$

where $w'_k \equiv W'_{kk}$, are then folded, marginalized, and scaled as before to yield $p(y^2|\mathbf{y}', \boldsymbol{W}')$. For the given signal $\mathbf{y}$, we first test the method for $\boldsymbol{V} = \boldsymbol{I}$ with trace N_d, with the results shown in Figure 13 (a) and (b). In this case $\mathbf{y}' = \mathbf{y}$, and $p(y^2|\mathbf{y}', \boldsymbol{W}')$ is the same as $p(y^2|\mathbf{y}, \sigma^2 = 1)$ from the previous paragraph. Now we suppose the data variance is given by a symmetric Toeplitz matrix $V_{jk} = 1/(1 + |j - k|)$ with the same trace N_d. The orthogonalized data $\mathbf{y}'$ shown in panel (c) have a distribution $p(y^2|\mathbf{y}', \boldsymbol{W}')$ shown in (d) with the same expectation value $\int y^2 p(y^2|\mathbf{y}', \boldsymbol{W}') dy^2$ as before. The orthogonalization of the data does not affect the expectation value of the normalized energy distribution but does affect its shape according to the independent datum distributions.

To make contact with our goal, we must now identify y^2 with the normalized spectral energy $E_{\mathbf{X}}/N_d\Delta_t$ given by the amplitude coefficients $\mathbf{A}$. The merit function $F_{RPQ} \equiv R + P + Q$ now includes a term $Q \equiv -\log p(E_{\mathbf{X}}|\mathbf{y}, \boldsymbol{V})$ for the distribution of the spectral energy and does not include any explicit constraint. The expression $E_{\mathbf{X}}$ is not a "hidden" variable but rather an auxiliary variable defined in terms of the model parameters. The term Q is in effect (the negative logarithm of) an additional prior factor in the amplitude space $\mathbf{A}$ which accounts for the distribution of the spectral energy given the data and its variance. Furthermore, the factors E_y and μ must now be rewritten in terms of $E_{\mathbf{X}}$ where they appear in the Poisson prior P of Equation (26). The implementation of these improvements is currently under investigation; we have outlined only one possible approach here, and some variation might be found to work better in practice.

8 Conclusions

In this article we have applied the principle of maximum entropy to power spectral density estimation using the one-sided Fourier transform in the context of discrete, irregular signal sampling. We have dispensed with the arbitrary weighting of the entropy term in the merit function in favor of a constraint on the spectral energy. The prior is rewritten in terms of the continuous Poisson distribution whose Stirling approximation gives the familiar entropy expression for unnormalized distributions. In the limit of vanishing errors, the spectrum with maximum evidence is equal to that with maximum likelihood given by the forward transform coefficients, and in the limit of extreme errors it approaches a flat power spectrum with the same energy as the signal. An outline of improvements to the method to obtain the variance of the spectral coefficients is also given.

As an example, we have evaluated the power spectrum of an irregularly sampled record of stellar luminosity for the star VCas. Several prominent peaks in the spectrum survive the smoothing action of the MaxEnt algorithm, which prevents the overestimation of structure in the spectrum when confronted with imperfect data. In that sense, the MaxEnt algorithm gives a more conservative estimate for the power spectrum than the forward transform. As actual measurements necessarily are accompanied by measurement errors, the incorporation of their effect through the principle of maximum entropy is suggested to those who use the Fourier transform for power spectral density estimation.

9 Appendix

In deriving the usual entropy expression, recourse is made to the discrete Poisson distribution

$$p_\mu^k = e^{-\mu}\mu^k/k! \propto \mu^k/k! \tag{44}$$

with parameter μ for integer $k \geq 0$, where the proportionality is given by the normalization $\sum_{k=0}^{\infty} \mu^k/k! = e^\mu$. For μ and k in some units except for the exponent, there are k unit factors in the numerator which cancel k unit factors in the denominator. While it has been suggested (Marsaglia 1986) that the cumulative distribution is what should be generalized to the continuum, it seems more intuitive that the probability density be generalized through the obvious substitution $k! \to \Gamma(n+1)$ for continuous $n \geq 0$, giving a continuous Poisson density $p_\mu^n \propto \mu^n/\Gamma(n+1)$. About the only objection one could raise for such a density function is one of normalization over an infinite axis, which itself is mooted by consideration of some finite cutoff $n_{\max}$ larger than any scale of interest in the problem at hand. The use of an unnormalizable distribution is implicit in the maximum likelihood method, as the uniform distribution is only finite when considered over a restricted domain.

As almost any non-negative function with finite domain can be normalized to a probability density, let us consider the normalization of the continuous Poisson distribution over an infinite axis. Can we show that $\int_0^\infty dn\, \mu^n/\Gamma(n+1) = e^\mu$? Taking the derivative with respect to the parameter μ of both sides, the RHS is the definition of the exponential function, $\partial_\mu e^\mu \equiv e^\mu$. The derivative of the LHS leads to the expression

$$\partial_\mu \int_0^\infty \frac{\mu^n\, dn}{\Gamma(n+1)} = \int_0^\infty \frac{n\mu^{n-1}\, dn}{\Gamma(n+1)} = \int_0^\infty \frac{\mu^{n-1}\, dn}{\Gamma(n)}\,, \tag{45}$$

whereupon shifting the limits down by one unit,

$$\int_{-1}^\infty \frac{\mu^n\, dn}{\Gamma(n+1)} = \int_{-1}^0 \frac{\mu^n\, dn}{\Gamma(n+1)} + \int_0^\infty \frac{\mu^n\, dn}{\Gamma(n+1)} \tag{46}$$

$$= \int_0^\infty \frac{\mu^n\, dn}{\Gamma(n+1)}\,, \tag{47}$$

where the last step follows from the observation that our original object, the probability density, is zero over the negative axis, $p_\mu^n = 0$ for $n < 0$. We hesitate to call this example a proof of the relation, as the most formal of mathematicians might object to the heuristic final step, but it is certainly highly suggestive that the normalization carries over to the continuum unaltered.

Acknowledgements We acknowledge with thanks the variable star observations from the AAVSO International Database contributed by observers worldwide and used in this research.

References

Boyd, J.P.: Journal of Computational Physics **103**(2), 243 (1992). doi:10.1016/0021-9991(92)90399-J

Bretthorst, G.L.: Bayesian Spectrum Analysis and Parameter Estimation. Springer, Berlin, Germany (1988)

Bryan, R.: European Biophysics Journal **18**, 165 (1990). doi:10.1007/BF02427376

Buck, B., Macaulay, V.A.: In: G., E., P., N., R., S.C. (eds.) Maximum Entropy and Bayesian Methods, Seattle 1991, p. 241. Kluwer Academic Publishers, Netherlands (1992)

Cenker, C., Feichtinger, H.G., Herrmann, M.: In: Computers and Communications, 1991. Conference Proceedings., Tenth Annual International Phoenix Conference on, p. 483 (1991). IEEE

D'Agostini, G.: ArXiv Physics e-prints (1998). Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting (Oxford Science Publications). arXiv:physics/9811046

Durrett, R.: The Essentials of Probability. Duxbury Press, A Division of Wadsworth, Inc., Belmont, CA (1994)

Gregory, P.: Bayesian Logical Data Analysis for the Physical Sciences. Cambridge University Press, New York, NY, USA (2005)

Henden, A.A.: 2011 Observations from the AAVSO International Database. private communication

Johnson, R.W.: Chapter 6. Extended Wavelet Transform for Discretely Sampled Data. In: del Valle, M., noz Guerrero, R.M., Salgado, J.M.G. (eds.) Wavelets: Classification, Theory and Applications, p. 125. Nova Science Publishers, Hauppauge, NY (2012). to appear

Lomb, N.R.: Astrophysics and Space Science **39**, 447 (1976). doi:10.1007/BF00648343

MacKay, D.J.C.: Neural Computation **11**(5), 1035 (1999)

Malik, W.Q., Khan, H.A., Edwards, D.J., Stevens, C.J.: In: Advances in Wired and Wireless Communication, 2005 IEEE/Sarnoff Symposium on, p. 125 (2005). IEEE

Marsaglia, G.: Computers and Mathematics with Applications **12**(5-6, Part 2), 1187 (1986). doi:10.1016/0898-1221(86)90242-7

Press, W., Teukolsky, S., Vetterling, W., Flannery, B.: Numerical Recipes in C, 2nd edn. Cambridge University Press, Cambridge, England (1992)

Scargle, J.D.: The Astrophysical Journal **263**, 835 (1982). doi:10.1086/160554

Sivia, D.S.: Data Analysis: A Bayesian Tutorial. Oxford Science Publications. Oxford University Press, Oxford, UK (1996)

Skilling, J.: In: Skilling, J. (ed.) Maximum Entropy and Bayesian Methods, Cambridge, England, 1988, p. 45. Kluwer Academic Publishers, Netherlands (1989)

Strauss, C., Wolpert, D., Wolf, D.: In: Mohammad-Djafari, A., Demoments, G. (eds.) Maximum Entropy and Bayesian Methods, Paris 1992, p. 113. Kluwer Academic Publishers, Netherlands (1993)

Wannier, G.H.: Statistical Physics. John Wiley and Sons, Inc., New York, NY (1969)

This manuscript was prepared with the AAS LaTeX macros v5.2.